Journal of Informetrics xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Correspondence

There should not be any mystery: A comment on sampling issues in bibliometrics

1. Introduction

A research unit wants to assess whether its publications tend to be more cited than academic publications in general. This research unit could be anything from a lonesome researcher to a national research council sponsoring thousands of researchers. The unit has access to the (inverted) percentile ranks of n of its publications: each publication has an associated real number between 0 and 100, which measures the percentage of publications in its reference set that receive *at least* as many citations as its own citation count.[1] For instance, if the percentile rank of one publication is 10%, it means that 90% of publications in its reference set are less cited and 10% are cited at least as much. Now, say that the unit takes the *mean* of its n percentile ranks and reaches a value below 50%. Can the unit confidently conclude that it produces research that tends to be more cited than most publications?

The article by Richard Williams and Bornmann (2016) proposes to answer this kind of question by relying on standard statistical procedures of significance testing and power analysis. I am deeply sympathetic to this proposal. I find, however, their exposition to be sometimes clouded in mystery. I suspect that many readers will therefore be unconvinced by their proposal. In this comment, I endeavor to make a clearer case for their general strategy. It should not be mysterious why this strategy is sound. By clarifying the case, I show that some technical decisions of Williams and Bornmann are mistakes (choosing a two-tailed instead of a one-tailed test), lead to less efficient estimates (choosing the t-statistic instead of simply relying on the mean) or are not as prudent as they should be (presuming a particular standard deviation). Before making these technical points, I start with a more conceptual issue: in the next section, I dispel some confusion regarding the notion of randomness in the presentation of the authors.

2. About some mysterious claims on randomness

Williams and Bornmann offer an argument for using inferential statistics even when we have access to *all* the relevant data on the publications of the research unit under study (Section 3.1 in their article). An argument is needed because it is counter-intuitive to use inferential statistics when the full ‘population’ seems already accessible. Isn’t inferential statistics about making inferences from sample to population? Why would we need its methods if all the observations are already at hand?

The general argument – a compelling one according to me – is that these observations are realizations of an underlying data generating process constitutive of the research unit. The goal is to learn properties of the data generating process. The set of observations to which we have access, although they are all the *actual* realizations of the process, do not constitute the set of all *possible* realizations. In consequence, we face the standard situation of having to infer from an accessible set of observations – what is normally called the sample – to a larger, inaccessible one – the population. Inferential statistics are thus pertinent.

Williams and Bornmann report this argument, but they then slip into talking about “the extent to which citations may have been influenced by random factors” (p. 9). They come up with a distinction between “random” factors and other, non-random factors. In this non-random category, we would have, primarily it seems, “the quality of the material in the papers” (p. 9). In the category of random factors, the examples given are: “how many people chose to read a particular issue of a journal or who happened to learn about a paper because somebody casually mentioned it to them” (pp. 9-10).

This distinction seems impossible to seriously uphold. How “casual” must a mention of a paper be to count as a “by chance” (p. 9) effect on citations? If most people read a given issue of a journal because they are regular readers of this journal, does the “how many people” factor become a non-random factor? And “the quality of the material” in a paper, why

[1] See http://ipscience-help.thomsonreuters.com/incitesLiveESI/ESIGroup/fieldBaselines/PercentilesBaselines.html.

http://dx.doi.org/10.1016/j.joi.2016.09.009

is this one not random? I personally cannot predict when I embark on a research project how important the output will be. Who does? The claim might instead be that a given quality translates into a number of citations in a deterministic fashion: if you can keep all the ‘chancy’ factors equal, an identical change in quality will be associated with a determinate change in the number of citations. This claim will sound odd to many. In contrast, I am inclined to endorse it. But I would add that, if it is plausible for the “quality” factor, it is also plausible for what the authors take to be “random” factors. For instance, if you keep all the other factors equal, a change in how many people read the paper will plausibly translate into citations.

Given how dubious the distinction between random and non-random factors is, it would be bad news if the strategy proposed by Williams and Bornmann required it. But there is no bad news: the mysterious distinction is irrelevant to the issue at stake. There is no need to be able to separate random from non-random factors to be justified in using inferential statistics even though all the realizations of the unit are given. What is needed is simply that the quantity *of interest* (here percentile ranking) is a random variable. And ‘random variable’ means that it is not possible in any given case to predict with certainty which value of the variable will be realized, although each value (or interval of values) has a probability of occurring. There is no need to track where the chance factors are among the factors causing the quantity.

In fact, random variables can even be the result of fully deterministic systems. Since Williams and Bornmann take the example of coin tosses as an analogy for their argument (p. 10), let me use the same case. Repeated tosses of a fair coin are properly modeled as independent realizations of a Bernouilli distribution with $p = .5$. The outcome is thus a random variable. Someone might want to say that coin tossing is affected by “chancy,” “random” factors. But the statistical modeling of the phenomenon is also fully compatible with a belief that the data generating process is perfectly deterministic (Ford, 1983): if only the *exact* initial conditions of the situation were known – angle of rotation, velocity, distance, etc. — and the *exact* laws of motion, the outcome could be perfectly predicted. Yet, the statistical modeling of the situation remains the best option because we have limited knowledge of the initial conditions and the laws of motion, and the outcome depends so crucially on these conditions (i.e., it is a system of deterministic chaos).

The same holds for the data generating processes responsible for percentile ranks. Statistical modeling would remain a relevant modeling choice even if it is believed that there is nothing that happens “by chance” here. The phenomenon is best modeled statistically because there is limited knowledge of the actual factors present in each specific situation and how these factors combine to produce the result. There is no need of a mysterious quest for the “random factors.” And there is room for inferential statistics: since realizations of percentile ranks follow a probability distribution, the full set of actual realizations (the sample) is not equivalent to the set of all possible realizations (the population). There is thus a gap that can be filled by inferential methods.

3. Why a specific *t*-test is fine

I now turn to discuss the technique proposed by Williams and Bornmann: a one sample *t*-test to decide whether the research unit in my introduction tends to produce research with more impact than most publications. It is a simple test, which is indeed appropriate here. Let me set up the problem and work through it to show the appropriateness of the solution. I will also argue that Williams and Bornmann turn out to use the wrong *t*-test and that, in fact, they could have done something even simpler.

Percentile ranks are attributed to each publication in a reference set by ordering these publications by their number of citations (from most cited to least cited) and mapping this ordering on the interval from 0 to 100. Given this procedure, the distribution of publications in the reference set is known: their distribution is approximately uniform on the interval [0,100], with mean $\mu_0 = 50$ and standard deviation $\sigma_0 = 28.87$.[2] If the output of the research unit is neither more nor less

[2] For a continuous variable following a perfectly uniform distribution on the interval [0,100], the value for the mean is obviously 50. The value for the variance σ_0^2 (you take the square root for the standard deviation) can be computed based on the probability density, which is simply $f(x) = 0.01$ for $x \in [0,100]$ and 0 otherwise:

$$\begin{aligned}\sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_0^{100} (x-50)^2 0.01 dx = 0.01 \int_0^{100} (x^2 - 100x + 2500)dx \\ &= 0.01\left(\frac{100^3}{3} - \frac{100 \times 100^2}{2} + 2500 \times 100\right) = \frac{100^2}{3} - \frac{100^2}{2} + 2500 = 2500 - \frac{10000}{6} = \frac{5000}{6} = 833.\bar{3} \approx 28.87^2\end{aligned}$$

As Williams and Bornmann note (fn. 1), the distribution of percentile ranks for the reference set is only *approximately* uniform because ties in citation counts result in publications sharing the same percentile rank. In particular, reference sets typically have a disproportionate amount of documents with no citation; all these documents have, according to the standard definition of percentile ranks, the same rank of 100. It is beyond the scope of this comment to systematically assess how closely the distribution of actual percentile ranks for a variety of reference sets tracks the uniform distribution. Yet, I have computed the means and standard deviations of five reference sets constructed from a corpus of economics documents indexed by Web of Science. Each reference set is made of documents for a single year between 1996 and 2000 inclusively. The sizes of the sets range from 7977 to 8356 documents. As expected, the actual means of percentile ranks lie slightly above the assumed value of 50, the maximal value being 52.7. The standard deviations also lie above the assumed 28.87, the maximal value being 31.2. These departures from the assumed values are arguably small. Given that there is evidence that reference sets in economics are further from the uniform distribution than reference sets in major disciplines such biology, physics and so on (Waltman & Schreiber, 2013, Table 1), my results about economics give some justification to the claim that the uniform distribution is a pretty close approximation to a large proportion of reference sets out there. I will thus proceed for the rest of this comment by assuming the uniform distribution at the level of the reference set.

Please cite this article in press as: Claveau, F. There should not be any mystery: A comment on sampling issues in bibliometrics. *Journal of Informetrics* (2016), http://dx.doi.org/10.1016/j.joi.2016.09.009